CR 134312

[redacted]

(NASA-CR-134312) OPTIMAL DESIGN OF AN UNSUPERVISED ADAPTIVE CLASSIFIER WITH UNKNOWN PRICES (Rice Univ.) 32 p HC $4.75 CSCL 12A

N74-28054

Unclas G3/19 43436

ICSA

INSTITUTE FOR COMPUTER SERVICES AND APPLICATIONS

RICE UNIVERSITY

32

Optimal Design of an Unsupervised Adaptive Classifier with Unknown Priors

by

Demetrios Kazakos
ICSA
Rice University

ABSTRACT

An adaptive detection scheme for M hypotheses is analyzed. We assume that the probability density function under each hypothesis is known, and that the prior probabilities of the M hypotheses are unknown and sequentially estimated. Each observation vector is classified using the current estimate of the prior probabilities. Using a set of nonlinear transformations, and applying stochastic approximation theory, we design an optimally converging adaptive detection and estimation scheme.

The optimality of the scheme lies in the fact that convergence to the true prior probabilities is ensured, and that the asymptotic error variance is minimum, for the class of nonlinear transformations considered.

We obtain also an expression for the asymptotic mean square error variance of our scheme.

Institute for Computer Services & Applications
Rice University
Houston, Texas 77001

May, 1974

Research supported under NASA Contract NAS 9-12776

I. Introduction

In general, there are three approaches to nonsupervised learning classifiers-detectors.

The first approach is the method of mixtures, in which we estimate the unknown parameters of a mixture distribution of input patterns. [1] It has been found, in general, that learning algorithms of this type are not simple to implement, and that they are slow to coverge, even though convergence to the true parameters can be guaranteed under certain loose restrictions. [2]

The second approach is the method of constructing a discriminant function by an iterative procedure. It is simpler, but usually it does not lead to optimal classification. One of the main techniques applicable to this approach is clustering, which has been investigated in [3] , [4] , and elsewhere.

The third approach is the decision directed method. It is a straightforward application of supervised learning methods, hence it is simple. Scudder [6] , Agrawala [5] , Davisson and Schwartz [7] , have discussed some learning algorithms based on this approach. The disadvantage of the method is that the estimates are usually asymptotically biased, due to classification errors.

In the present paper, we will use an improved version of the decision-directed approach.

A decision-directed detector (classifier) uses previous decisions to estimate unknown parameters. On the basis of these estimates, the detector structure is modified for subsequent decision. The fundamental idea is that the detector assumes all past decisions correct, and on this basis, he tries to improve his performance by adjusting his decision

parameters. Applying this idea, Scudder [6] considered the binary detection of unknown signal versus no signal in noise. He estimated the signal as the sample mean estimate based on observations that were classified as containing the signal. Convergence of his estimate was heuristically argued. His estimate was asymptotically biased.

Davisson and Schwartz [7] studied a decision-directed detector using previous decisions to estimate the prior probabilities as relative frequencies of decisions in favor of each hypothesis. Their estimate is asymptotically biased, but for certain applications the bias is sufficiently small for practical purposes.

They found bounds to the probability that the estimate of the prior will "run away" to 0 or 1.

In an unpublished study [8], the author has found similar bounds to the "run away" probability, for a decision-directed receiver where both the signal amplitude and the prior probability are sequentially updated, using previous decisions.

In order to have optimal performance of an M-ary detection scheme, accurate knowledge of the prior probabilities is necessary.

The degradation in the probability of error when in correct prior probabilities are used, has been computed in a closed form.

In the present paper, the method of Davisson and Schwartz [7] for simultaneous detection and sequential estimation of the prior probabilities, is generalized to arbitrary probability density functions and substantially improved by the introduction of nonlinear transformations.

The prior probability updating procedure is made to converge to the true values, in an optimal fashion, in the sense that the asymptotic error variance is minimized.

A stochastic approximation theorem due to Sacks [9] is invoked.

Possibly, the most interesting conclusion of the present work, is the fact that we can improve the behavior of a sequential estimation procedure by using a memoryless nonlinear transformation.

II. Binary detection with inaccurate prior probability

We now consider the problem of binary detection with inaccurate knowledge of the prior probability.

Let H_1, H_2 be the two hypotheses, with corresponding prior probabilities π and $1 - \pi$, and probability density functions $f_1(x)$ and $f_2(x)$. The observation is x, $x \in E^n$. We pose the following convenient restriction on the p.d.f's.

$\forall x \in E^n$, $f_1(x) > 0$, $f_2(x) > 0$ and continuous.

It is well known that the optimal decision rule that minimizes the average probability of error is :

Decide H_1 if $\pi f_1(x) \geq (1-\pi) f_2(x)$

Decide H_2 otherwise.

Hence, the knowledge of π is essential for optimality.

Let $Pe(p)$ be the average probability of error, when p is the estimate of π used in the decision rule.

It is then easy to show that :

$$Pe(p) - Pe(\pi) = \int_{R(\pi, p)} \left[\pi f_1(x) - (1-\pi) f_2(x) \right] dx \geq 0$$

where

$$R(\pi, p) = \left\{ x ; \min\left(\frac{1-\pi}{\pi}, \frac{1-p}{p} \right) \leq \frac{f_1(x)}{f_2(x)} \leq \max\left(\frac{1-\pi}{\pi}, \frac{1-p}{p} \right) \right\}$$

The above formula gives us in compact form the suboptimality of a decision rule that uses an incorrect estimate, p, of the true prior, π.

III. Adaptive detection-estimation scheme for 2 hypotheses

We now assume that $f_1(x)$, $f_2(x)$ are known and positive for all $x \in E^n$, and that π is unknown.

The method of simultaneous detection and estimation of π employed in [7] is the following.

Assume that n past observations $x_1 \ldots x_n$ have been classified, n_1 of them to H_1 and n_2 of them to H_2 . A natural estimate of π , is then

$$p_n = \frac{n_1}{n}$$

When observation x_{n+1} is received, its classification is then based on p_n :

$$\text{Decide } x_{n+1} \in H_1 \text{ if } \frac{f_1(x_{n+1})}{f_2(x_{n+1})} \geq \frac{1 - p_n}{p_n}$$

$$\text{Decide } x_{n+1} \in H_2 \text{ otherwise .}$$

p_n can be expressed as :

$$p_n = \frac{1}{n} \sum_{j=1}^{n} W_j$$

where

$$W_{k+1} = \begin{cases} 1 & \text{if } \dfrac{f_1(x_{k+1})}{f_2(x_{k+1})} \geq \dfrac{1 - p_k}{p_k} \\ 0 & \text{otherwise .} \end{cases}$$

Written in a recursive estimation form :

$$p_{n+1} = p_n - \frac{1}{n+1} \left[p_n - W_{n+1} \right]$$

The expected value of p_{n+1} conditioned on p_n is given by :

$$E(p_{n+1} \mid p_n) = p_n - \frac{1}{n+1} \left[p_n - E(W_{n+1} \mid p_n) \right]$$

From the above form, it is clear that if p_n converges to a value q, then this value will satisfy the equation :

$$q = E(W_{n+1} \mid p_n = q)$$

or

$$\pi \int_{R(q)} \left[f_1(x) - f_2(x) \right] dx + \int_{R(q)} f_2(x)\, dx - q = 0$$

where

$$R(q) = \left\{ x ; \frac{f_1(x)}{f_2(x)} \geq \frac{1-q}{q} \right\}$$

In general, the root of this equation is not equal to π, and therefore, the procedure leads to an asymptotically biased estimate of π.

We now introduce a modified sequential detection-estimation scheme, in order to improve on the original.

Let $L(x)$, $g(x)$ be two nonlinear functions defined for $x \in [0, 1]$. Then, the following estimation algorithm is proposed :

$$p_{n+1} = p_n - \frac{1}{n+1} L(p_n) \cdot \left[g(p_n) - W_{n+1} \right]$$

The new regression function of the modified algorithm, is :

$$M(p_n) = E\left[L(p_n) \left[g(p_n) - W_{n+1} \right] \mid p_n \right]$$

$$= L(p_n) \left\{ g(p_n) - \int_{R(p_n)} \left[\pi f_1(x) + (1-\pi) f_2(x) \right] dx \right\}$$

Necessary condition for having an asymptotically unbiased estimate, is to have the value $p_n = \pi$ as a root of the regression equation :

$$M(\pi) = 0$$

This condition is achieved by choosing the function g as follows :

$$g(s) = \int_{R(s)} \left[s\, f_1(x) + (1 - s)\, f_2(x) \right] dx$$

for $s \in [0 , 1]$.

Substituting g into the previous equation, we have :

$$M(p_n) = L(p_n) \cdot (p_n - \pi) \cdot G(p_n)$$

where

$$G(s) = \int_{R(s)} \left[f_1(x) - f_2(x) \right] dx$$

The function $G(s)$ is monotone increasing for $s \in [0 , 0.5]$ and monotone decreasing for $s \in [0.5 , 1]$, as shown in Appendix I . Also, $G(0) = G(1) = 0$, $G(s) > 0$ for $s \in (0 , 1)$.
The form of the function $G(s)$ is given in Fig. 1.

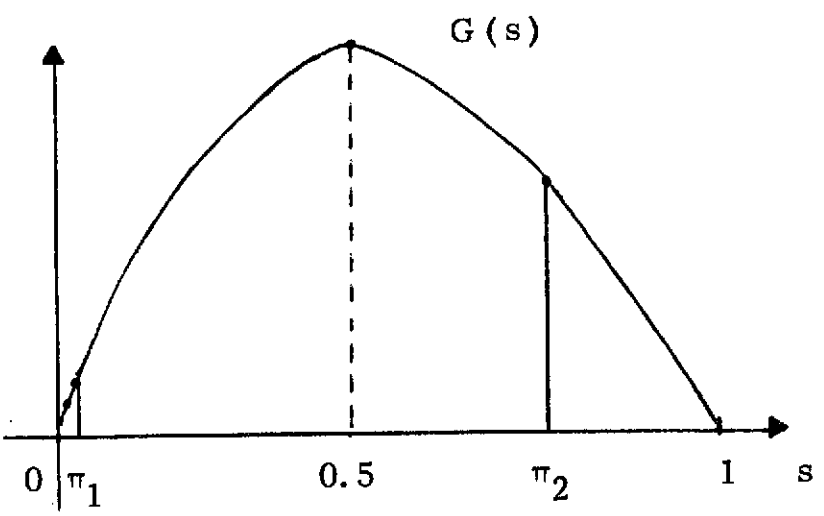

Fig. 1

For reasons to be explained, we assume that we have the knowledge that the unknown prior π , lies between π_1 and π_2 , where $0 < \pi_1 < 0.5$, $0.5 < \pi_2 < 1$.

Let

$$Z(p) = L(p)\left[g(p) - W \right] - M(p)$$

for

$$p \in I_2 ,$$

where

$$I_2 = \left[\pi_1 , \pi_2 \right]$$

and

$$W = \begin{cases} 1 & \text{if } \dfrac{f_1(x)}{f_2(x)} \geq \dfrac{1-p}{p} \\ 0 & \text{otherwise .} \end{cases}$$

Then, the Robbins-Monro Stochastic Approximation procedure that gives the sequential estimates of π, is written :

$$p_{n+1} = p_n - (n+1)^{-1} \left[M(p_n) + Z(p_n) \right]$$

We invoke now a theorem due to Sacks [9] , in a slightly modified version, to fit the circumstances.

The assumptions to be checked are :

Assumption (1)

$(x - \pi) M(x) > 0$

for all $x \in I_2$, $x \neq \pi$

Assumption (2)

For all $x \in I_2$ and some positive constant k_1 ,

$\left| M(x) \right| \leq k_1 \left| x - \pi \right|$

and for every t_1 , t_2 , such that $0 < t_1 < t_2 < \infty$

$\inf \left| M(x) \right| > 0$ where the inf is taken for $x \in I_2$ and $t_1 \leq |x - \pi| \leq t_2$

Assumption (3)

For all $x \in I_2$,

$M(x) = a_1 (x - \pi) + \delta (x , \pi)$

where $\delta (x , \pi) = 0 \left(|x - \pi| \right)$ as $|x - \pi| \longrightarrow 0$

and where $a_1 > 0$

Assumption (4)

(a) $\sup_{x \in I_2} E Z^2 (x) < \infty$

(b) $\lim_{x \to \pi} E Z^2 (x) = \sigma^2$

Assumption (5)

$\{ Z (x) \}$ are identically distributed random variables (conditioned on x).

Assume, further, that $a_1 > \frac{1}{2}$.

Then, the stochastic approximation procedure

$$p'_{n+1} = p_n - (n+1)^{-1} \cdot L(p_n) \cdot \left[g(p_n) - W_{n+1} \right]$$

$$p_{n+1} = \begin{cases} \pi_1 & \text{if } p'_{n+1} \leq \pi_1 \\ p'_{n+1} & \text{if } \pi_1 \leq p'_{n+1} \leq \pi_2 \\ \pi_2 & \text{if } \pi_2 \leq p'_{n+1} \end{cases}$$

converges to π, and the error $n^{\frac{1}{2}} (p_n - \pi)$ is asymptotically normally distributed with mean 0 and variance $\sigma^2 (2a_1 - 1)^{-1}$

Note :

If the condition $a_1 > \frac{1}{2}$ is not satisfied, the convergence of the error variance is slower than n^{-1}, as Sakrison [10] points out.

For our particular case,

$$M(p_n) = L(p_n) \cdot (p_n - \pi) \cdot G(p_n)$$

hence, if we restrict $L(x) > 0$, for $x \in I_2$,

the assumption (1) is satisfied.

Also, if we further restrict $L(x)$ to be bounded, assumption (2) is easily satsified.

That is, the condition :

$$0 < K_2 < L(x) < K_3, \quad \text{for } x \in I_2,$$

satisfies assumptions (1) and (2). Assumption (3) is satisfied if we let

$$a_1 = M'(\pi).$$

For our case, if we assume that $G'(x)$ exists, we have :

$$M'(x) = L'(x)(x - \pi) \cdot G(x) + L(x) \cdot G(x) + L(x) \cdot G'(x)$$

$$M'(\pi) = L(\pi)\left[G(\pi) + G'(\pi)\right]$$

Also,

$$\sigma^2 = \lim_{Z \to \pi} E\left\{ \left[L(x)\left[g(z) - W\right] - M(Z)\right]^2 \mid Z \right\}$$

After some straightforward calculations, we find :

$$\sigma^2 = L^2(\pi)\, g(\pi)\left(1 - g(\pi)\right)$$

The formula for the asymptotic error variance then becomes :

$$\text{var } n^{\frac{1}{2}} (p_n - \pi) \longrightarrow \frac{L^2(\pi)\, g(\pi) \left(1 - g(\pi) \right)}{2L(\pi) \left[G(\pi) + G'(\pi) \right] - 1}$$

In Fig. 2, we plot the function $G(\pi) + G'(\pi)$.

For $0 \leq \pi < \pi_o$, the function is positive, and for $\pi \geq \pi_o$, it is negative.

The crossover point π_o , lies between 0.5 and 1 .

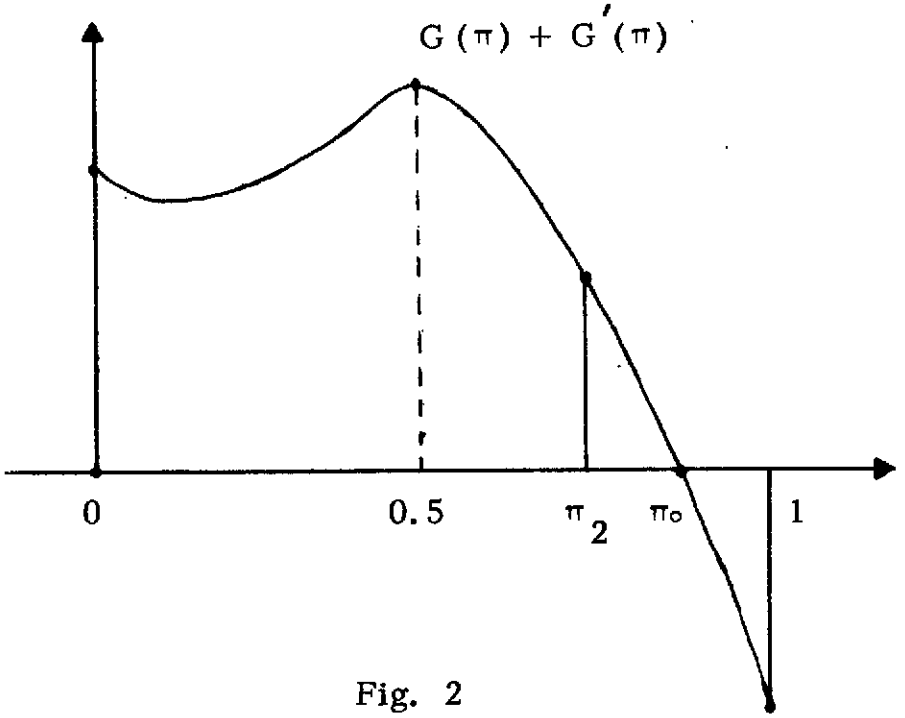

Fig. 2

Therefore, if $\pi_2 < \pi_o$, we are ensured that $G(\pi) + G'(\pi) > 0$

$\forall\ \pi \in I_2$.

We are now in a position to find the optimal function $L(\pi)$, that will minimize the variance expression. It is a straightforward matter to see that the optimum function L is :

$$L(\pi) = \left[G(\pi) + G'(\pi) \right]^{-1} , \quad \pi \in I_2$$

Using the above L, the achievable minimum asymptotic error variance is :

$$V = \frac{g(\pi)\left(1-g(\pi)\right)}{\left[G(\pi) + G'(\pi) \right]^2} , \quad \pi \in I_2$$

In Appendix I, we prove why $G(Z)$ has the form of Fig. 1 for general density functions.

In Appendix II, it is shown how we can easily compute $g(\pi)$, $G(\pi)$, $G'(\pi)$ for the multivariate Gaussian case.

IV. Adaptive M-ary detection-estimation scheme with unknown priors

In the present section, we are considering the problem of adaptive detection with unknown prior probabilities under M hypotheses, $M > 2$. The hypotheses $H_1 \ldots H_M$ correspond to prior probabilities $\pi_1 \ldots \pi_M$, and probability density functions $f_1(x)$, $f_2(x)$, ..., $f_M(x)$.

For simplicity, we assume again that $f_i(x)$ are positive and continuous $\forall x \in E^n$

We need to estimate $M - 1$ of the prior probabilities only.

Let p_n^j, $j = 1, \ldots, M$ be the n^{th} estimate of π_j

Let $\pi = \left[\pi_1 , \pi_2 \cdots \pi_{M-1} \right]^T$

and $P_n = \left[p_n^1 \; p_n^2 \cdots p_n^{M-1} \right]^T$

Then, the observation x_{n+1} will be classified according to the decision rule :

$$\text{Decide } H_k \text{ if } p_n^k f_k(x_{n+1}) = \max_m p_n^m f_m(x_{n+1})$$

Again, to achieve optimality in our decision rule, we must find a method of forcing the vector p_n to converge to the true value π, "as fast as possible."

As before, a natural estimate of π is the relative frequencies vector decisions in favor of the several classes.

Let n be the number of past observations, and let n_k, $k = 1, \ldots, M$ be the number of those classified as belonging to H_k.

$$n = n_1 + n_2 + \ldots . n_M .$$

Then, a natural estimate of π_k is :

$$p_n^k = \frac{1}{n} \sum_{s=1}^{n} W_s^k$$

where

$$W_{s+1}^k = \begin{cases} 1 & \text{if } p_n^k f_k(x_{s+1}) = \max_m p_n^m f_m(x_{s+1}) \\ 0 & \text{otherwise.} \end{cases}$$

$k = 1, 2, \ldots, M.$

We have the following sequential estimation algorithm :

$$p_{n+1}^k = p_n^k - (n+1)^{-1} \left(p_n^k - W_{n+1}^k \right)$$

$k = 1, \ldots, M.$

Written in a vector form :

$$P_{n+1} = P_n - (n+1)^{-1} \cdot \begin{bmatrix} p_n^1 - W_{n+1}^1 \\ p_n^{M-1} - W_{n+1}^{M-1} \end{bmatrix}$$

This algorithm has the fault of the original one in one dimension.

It is asymptotically biased.

The modification to be introduced now is a nontrivial generalization of the one dimensional case.

For reasons of convergence, we assume we have the knowledge that $\pi \in I_M$, where I_M is a subset of $(0, 1)^{M-1}$, to be specified later.

In the case $M = 2$, we have the previously defined internal $I_2 = [\pi_1, \pi_2]$.

A desirable property that we will assign to the internal I_M, is :

If

$$\pi \in I_M, \quad \pi_i \in [\ell_i, h_i]$$

$$i = 1, \ldots, M - 1$$

where

$$0 < \ell_i < h_i < 1$$

ℓ_i are small positive numbers, and h_i are close to and below 1.

We will assume, therefore, that we have knowledge of the region I_M, and hence ℓ_i, h_i are known.

We define the saturation function,

$$\text{sat}[x, \ell, h] = \begin{cases} \ell & \text{if } x \leq \ell \\ x & \text{if } x \in [\ell, h] \\ h & \text{if } x \geq h \end{cases}$$

After that, we are describing the modified sequential estimation procedure for the vector π.

$$P'_{n+1} = P_n - (n+1)^{-1} \cdot A(P_n) \cdot L(P_n) \cdot \begin{bmatrix} g_1(P_n) - W^1_{n+1} \\ g_2(P_n) - W^2_{n+1} \\ \vdots \\ g_{M-1}(P_n) - W^{M-1}_{n+1} \end{bmatrix}$$

and

$$p^j_{n+1} = \text{sat}\left[p'^{,j}_{n+1}, \ell_j, h_j\right]$$

$$j = 1, \ldots, M - 1$$

where

W^{j}_{n+1} are the previously defined random variables

$g_j(P_n)$ are scalar functions, defined for $P_n \in I_M$

$A(P_n)$ is a positive scalar function defined on I_M

and

$L(P_n)$ is an $(M-1) \times (M-1)$ matrix

$$L(P_n) = \left\{ L_{ij}(P_n) \right\} \quad , \quad i, j = 1, \ldots, M-1$$

where

$L_{ij}(P_n)$ are scalar functions defined on I_M .

All of the above functions L , g , A will be designed for improvement of the original algorithm.

We compute :

$$E\left[g_j(P_n) - W^{j}_{n+1} \mid P_n \right]$$

$$= g_j(P_n) - \Pr\left[p^{j}_{n} f_j(x) = \max_k p^{k}_{n} f_k(x) \right]$$

$$= g_j(P_n) - \sum_{s=1}^{M} \pi_s \int_{R_j(P_n)} f_s(x)\, dx$$

where

$$R_j(P_n) = \left\{ x ; \; p^{j}_{n} f_j(x) = \max_k p^{k}_{n} f_k(x) \right\}$$

If we let

$$g_j(P_n) = \sum_{s=1}^{M} p^{s}_{n} \int_{R_j(P_n)} f_s(x)\, dx$$

$$j = 1, \ldots, M-1$$

then

$$E\left[g_j(P_n) - W^j_{n+1} \,\middle|\, P_n \right]$$

$$= \sum_{s=1}^{M} \left(p_n^s - \pi_s \right) \int_{R_j(P_n)} f_s(x)\, dx$$

$$= \sum_{s=1}^{M-1} \left(p_n^s - \pi_s \right) \int_{R_j(P_n)} f_s(x)\, dx + \left(p_n^M - \pi_M \right) \int_{R_j(P_n)} f_M(x)\, dx$$

$$= \sum_{s=1}^{M-1} \left(p_n^s - \pi_s \right) \int_{R_j(P_n)} \left[f_s(x) - f_M(x) \right] dx$$

We defind the following $(M - 1) \times (M - 1)$ matrix $F(P_n)$, with elements

$$F_{sj}(P_n) = \int_{R_j(P_n)} \left[f_s(x) - f_M(x) \right] dx$$

$$s, j = 1, \ldots, M - 1$$

Then, the $(M - 1)$ dimensional regression function for the modified multidimensional algorithm, is :

$$M(P_n) = A(P_n) \cdot L(P_n) \cdot E\left[g_1(P_n) - W^1_{n+1}, \right.$$

$$\left. g_2(P_n) - W^2_{n+1}, \ldots, g_{M-1}(P_n) - W^{M-1}_{n+1} \,\middle|\, P_n \right]^T$$

or

$$M(P_n) = A(P_n) \cdot L(P_n) \cdot F(P_n) \cdot (P_n - \pi)$$

Hence :

$$M(\pi) = 0$$

We note that by properly choosing the functions $g_j(P_n)$, we have achieved to make the procedure asymptotically unbiased, assuming that it converges. We still have to ensure convergence

Let

$$Z(P_n) = A(P_n)\ L(P_n)\ \left\{ \left[g_1(P_n) - W_{n+1}^1, \cdots, g_{M-1}(P_n) - W_{n+1}^{M-1} \right]^T - F(P_n)\ (P_n - \pi) \right\}$$

defined for $P_n \in I_M$.

To ensure convergence and asymptotic normality of the error, we now invoke the multidimensional version of Sacks theorem [9].

The assumptions to be satisfied, are :

Assumption (1)

$M(\pi) = 0$, and for every $e > 0$

$\inf (x - \pi)^T\ M(x) > 0$

where the inf is taken for

$x \in I_M$ and $e^{-1} > \| x - \pi \| > e$

Assumption (2)

There exists a positive constant k_1 such that :

$$\| M(x) \| \leq k_1 \| x - \pi \|$$

for all $x \in I_M$

Assumption (3)

For all $x \in I_M$,

$M(x) = B \cdot (x - \pi) + \delta(x, \pi)$

where B is an $(M - 1) \times (M - 1)$ positive definite matrix,

$\| \delta(x, \pi) \| = 0(\| x - \pi \|)$, as $\| x - \pi \| \longrightarrow 0$

Assumption (4)

$$\sup_{x \in I_M} E \left\| Z(x) \right\|^2 > \infty$$

$$\lim_{x \to \pi} E \; Z(x) \; Z^T(x) = S$$

where S is a nonnegative definite matrix.

Assumption (5)

$\{ Z(x) \}$ are identically distributed.

(conditioned on x)

Let $b_1 \ldots b_{M-1}$ be the eigenvalues of B in decreasing order.

Write $B = PDP^{-1}$, where P is orthogonal and D is the diagonal matrix with diagonal elements $(b_1 \ldots b_{M-1})$

Let s_{ij} be the (i, j)th element of S, and let s^*_{ij} be the (i, j)th element of $S^* = P^{-1} S P$

Let $b_{M-1} > \frac{1}{2}$

Then, $n^{\frac{1}{2}} (P_n - \pi)$ is asymptotically normal, with zero mean, and covariance matrix PQP^{-1}, where Q is the matrix whose (i, j)th element is $(b_i + b_j - 1)^{-1} \cdot s^*_{ij}$

We now assume that the matrix $F(P)$ is nonsingular for all $P \in I_M$.

Satisfaction of the above condition depends on the form of the probability density functions

$f_1(x), \ldots, f_M(x)$, and on the interval I_M.

For reasons to be seen immediately, we choose :

$$L(P) = F^T(P)$$

Also, we restrict the scalar function $A(P)$ to be bounded :

$$0 < k_2 < A(P) < k_3 , \quad \forall P \in I_M$$

Hence, the inf of Assumption (1) is bounded from below :

$$\inf (x - \pi)^T M(x) = \inf (x - \pi)^T A(x) . F^T(x) F(x) (x - \pi) \geq$$

$$\geq k_2 e^2 \inf_{y \in I_M} p^2(y)$$

where $p^2(y)$ is the minimum eigenvalue of $F^T(y) F(y)$

Because of the assumption that $F(y)$ is nonsingular, $p^2(y)$ is positive. Hence, Assumption (1) is satisfied. Assumptions (2) and (4) are easily shown to hold, and Assumption (3) also holds, with

$$B = A(\pi) . F^T(\pi) F(\pi)$$

We need now to compute the covariance matrix S.

In Appendix III, we show that

$$S(\pi) = A^2(\pi) . F^T(\pi) . R(\pi) . F(\pi)$$

where $R(\pi)$ is a $(M - 1) \times (M - 1)$ matrix with elements

$$R_{km}(\pi) = g_k(\pi) \left(\delta_{km} - g_m(\pi) \right)$$

where

$$\delta_{km} = \begin{cases} 1 & \text{for } k = m \\ 0 & \text{for } k \neq m \end{cases}$$

If $q_1 \ldots q_{M-1}$ are the eigenvalues of $F^T F$ in decreasing magnitude, we can write

$$F^T F = P \, \mathrm{diag}(q_1 \ldots q_{M-1}) \cdot P^{-1}$$

where P is orthogonal.

The eigenvalues of B are then

$$b_i = A q_i$$

and

$$B = A \, P \, \mathrm{diag}(q_1 \ldots q_{M-1}) \, P^{-1}$$

$$S^* = A^2 \, P^{-1} \, F^T \, R \, F \, P$$

Let m_{ij} be the elements of the matrix

$$P^{-1} \, F^T \, R \, F \, P$$

Then

$$s^*_{ij} = A^2 \, m_{ij}$$

and Q has elements

$$A^2 \, (A q_i + A q_j - 1)^{-1} \cdot m_{ij}$$

To conclude :

For every $\pi \in I_M$, the matrices F, P, R are fixed. The only parameter to be adjusted, is $A(\pi)$.

The restrictions $A(\pi)$ must satisfy are :

$$0 < k_2 < A(\pi) < k_3 \qquad \forall \pi \in I_M$$

Furthermore, we must have

$$A(\pi) \cdot \min_k q_k (\pi) > \tfrac{1}{2}$$

This last condition is essential in order to have mean square convergence of the error of the order n^{-1}. If it is not satisfied, as Sakrison [10] points out, convergence is slower than n^{-1}.

The optimal choice of $A(\pi)$, hence, is the value that minimizes the trace of PQP^{-1}, where Q has elements

$$Q_{ij} = A^2 (A q_i + A q_j - 1)^{-1} \cdot m_{ij}$$

under the constraint :

$$A \min_k q_k > \tfrac{1}{2}$$

We have

$$T(A) = \text{trace}(P\,Q\,P^{-1}) = \text{trace}(Q\,P^{-1}\,P) = \text{trace}(Q)$$

Hence, we wish to minimize

$$T(A) = A^2 \sum_{k=1}^{M-1} m_{kk} (2 A q_k - 1)^{-1}$$

under the constraint :

$$2 A q_k > 1$$

for $k = 1, \ldots, M - 1$

where

$$m_{kk} > 0$$

and

$$q_1 > q_2 > \ldots > q_{M-1} > 0$$

The function $T(A)$ is positive in the region

$$A > (2 q_{M-1})^{-1}$$

For $A \longrightarrow +\infty$, $T(A) \longrightarrow +\infty$

and for $A \searrow (2 q_{M-1})^{-1}$, $T(A) \longrightarrow +\infty$

Since it is also a ratio of polynomials, it must have a number of local minima for $A > (2 q_{M-1})^{-1}$

The derivative of $T(A)$ is :

$$T'(A) = 2A \sum_{k=1}^{M-1} m_{kk} \frac{Aq_k - 1}{(2Aq_k - 1)^2}$$

For $A \geq (q_{M-1})^{-1}$, $T'(A) > 0$

Hence, the region of interest for seeking zeros of $T'(A)$, is

$$\left((2q_{M-1})^{-1} , (q_{M-1})^{-1} \right) = I(q_{M-1})$$

The number of zeros of $T'(A)$ is at most $2(M-1)$.

Let A_i , $i = 1 , \ldots , 2(M-1)$, be the zeros of $T'(A)$ that are in $I(q_{M-1})$.

Then, the optimal value of A is :

$$A_o(\pi) = \arg \left[\min T(A_i) \right]$$

The above procedure was done for a fixed $\pi \in I_M$.

Doing the same thing for a mesh in I_M , we can construct the optimal nonlinearity $A_*(\pi)$, $\pi \in I_M$.

Therefore, the trace of the error covariance matrix, has the asymptotic minimum value of :

$$\text{trace} \left[n^{\frac{1}{2}} E(p_n - \pi)(p_n - \pi)^T \right]$$

$$\longrightarrow A_o^2(\pi) \cdot \sum_{k=1}^{M-1} m_{kk}(\pi) \left[2A_o(\pi) q_k(\pi) - 1 \right]^{-1}$$

V. Conclusions

It has been shown that we can estimate efficiently the prior probabilities, even in the presence of detection errors. The cost we have to pay, is the construction of the above nonlinear functions of π.

For the binary case and for multivariate Gaussian probability density functions, the nonlinear transformation functions are easy to construct. The important conclusion is, that the use of nonlinear transformations can improve the properties of stochastic approximation methods.

APPENDIX I

For the two hypotheses and for general probability density functions, the function

$$G(z) = \int \left[f_1(x) - f_2(x) \right] dx$$

$$\frac{f_1(\xi)}{f_2(\xi)} \geq \frac{1 - z}{z}$$

has the form given in Fig. 1.

Proof :

a). For $z \in [0, 0.5]$,

$f_1(x) \cdot f_2^{-1}(x) \geq (1-z) \cdot z^{-1} \geq 1$

and $(1 - z)\, z^{-1}$ is monotone decreasing, hence $G(z)$ is monotone increasing. Also, $G(0) = 0$.

b). For $z \in [0.5, 1]$, $z^{-1}(1 - z) \leq 1$

$$G(z) = \int \left[f_1(x) - f_2(x) \right] dx - \int \left[f_2(x) - f_1(x) \right] dx$$

$$\frac{f_1(\xi)}{f_2(\xi)} \geq 1 \qquad 1 \geq \frac{f_1(\xi)}{f_2(\xi)} \geq \frac{1 - z}{z}$$

The second integral is a positive, monotone increasing function of z, hence $G(z)$ is monotone decreasing for $z \in [0.5, 1]$.

Also, $G(1) = 0$

Hence, $G(z)$ has a maximum at $z = 0.5$ and has the form given in Fig. 1 .

APPENDIX II

For the two hypotheses case and for Gaussian n-dimensional probability density functions, the nonlinearities $g(\pi)$, $G(\pi)$, $L(\pi)$ can be constructed by using a method due to Fukunaga and Krile [11] .
The essential characteristic of the method is the linear transformation of the observation vector to a new one that has components statistically independent under both hypotheses.
We are interested in computing the following two integrals :

$$S_1(\pi) = \int f_1(x)\, dx$$

$$\frac{f_1(\xi)}{f_2(\xi)} \geq \frac{1-\pi}{\pi}$$

$$S_2(\pi) = \int f_2(x)\, dx$$

$$\frac{f_1(\xi)}{f_2(\xi)} \geq \frac{1-\pi}{\pi}$$

for $\pi \in I_2$

Let $f_1(x) = N(x, 0, R_1)$

$f_2(x) = N(x, M, R_2)$

where $M = M_2 - M_1$ = difference of mean vectors.

Let A be the $n \times n$ matrix satisfying the relations :

$$A R_1 A^T = I$$

$$A R_2 A^T = \Lambda$$

where $\Lambda = \mathrm{diag}(\lambda_1 \ldots \lambda_n)$, with $\lambda_1 \ldots \lambda_n$ the eigenvalues satisfying :

$$\left| R_2 - \lambda R_1 \right| = 0$$

For R_1 , R_2 positive definite, all λ_i are positive.

Let $u = AM = (u_1 \ldots u_n)^T$

If x is the original observation vector, the new one is

$$Y = Ax = (y_1 \ldots y_n)^T$$

The statistics of Y are :

$$E(Y \mid H_1) = 0 , \quad E(Y \mid H_2) = \mu$$

$$E(YY^T \mid H_1) = I , \quad E\left[\left(Y - \mu\right) \left(Y - \mu\right)^T \middle| H_2 \right] = \Lambda$$

We can see easily, then, that

$$S_1(\pi) = Pr\left[U(Y) < 0 \mid H_1 \right]$$

where

$$U(Y) = \sum_{j=1}^{n} \left\{ y_j^2 - \frac{1}{\lambda_j} (y_j - \mu_j)^2 - \mathrm{Ln}\lambda_j \right\} - 2\mathrm{Ln}\left[\pi (1-\pi)^{-1} \right]$$

Since y_j are independent Gaussian random variables with zero mean and unit variance, the characteristic function of $U(Y)$ can be easily computed.

$$M_1(j\omega) = E\left\{ \exp\left(j\omega \cdot U(Y) \right) \middle| H_1 \right\}$$

$$= K(j\omega) \cdot \prod_{m=1}^{n} F_{1m}(j\omega)$$

where

$$K(j\omega) = \exp\left[-2j\omega \operatorname{Ln}\left[\pi(1-\pi)^{-1}\right]\right]$$

$$F_{1m}(j\omega) = \left(1 - j2a_{1m}\omega\right)^{-\frac{1}{2}} \cdot \exp\left[\frac{-(2a_{1m}\, b_{1m}\,\omega)^2}{2(1 - j2a_{1m}\omega)} - j\omega h_{1m}\right]$$

and

$$a_{1m} = 1 - \frac{1}{\lambda_m}$$

$$b_{1m} = \frac{u_m}{\lambda_m - 1}$$

$$h_{1m} = \frac{\left(a_{1m}\, b_{1m}\right)^2}{1 - a_{1m}} + \operatorname{Ln}\lambda_m$$

Also, if we make the transformation:

$$Z = \Lambda^{-\frac{1}{2}} A(x - M) = (z_1 \ldots z_n)^T$$

the vector Z has zero mean and unit variance Gaussian independent components.

Then,

$$S_2(\pi) = Pr\left[V(Z) < 0 \mid H_2\right]$$

where

$$V(Z) = \sum_{k=1}^{n}\left[(\lambda_k - 1)\left(Z_k + \frac{\sqrt{\lambda_k}\, u_k}{\lambda_k - 1}\right)^2 - \left(\frac{\mu_k^2}{\lambda_k - 1} + \operatorname{Ln}\lambda_k\right)\right] - 2\operatorname{Ln}\left[\pi(1-\pi)^{-1}\right]$$

The characteristic function of $V(Z)$, is then :

$$M_2(j\omega) = K(j\omega) \prod_{m-1}^{n} F_{2m}(j\omega)$$

where, $F_{2m}(j\omega)$ have the same form with $F_{1m}(j\omega)$, with corresponding parameters :

$$a_{2m} = \lambda_m - 1$$

$$b_{2m} = \sqrt{\lambda_m \mu_m} \left(\lambda_m - 1\right)^{-1}$$

$$h_{2m} = -\left(a_{2m}\ b_{2m}\right)^2 \left(1 + a_{2m}\right)^{-1} + \mathrm{Ln}\lambda_m$$

Hence, $M_1(j\omega)$ and $M_2(j\omega)$ can be easily expressed in terms of μ and λ.

Then, the functions $S_k(\pi)$, $k = 1, 2$ can be expressed as an integral involving $M_k(j\omega)$, as follows :

$$S_k(\pi) = 2^{-1} - p^{-1} \int_0^{+\infty} \omega^{-1} \mathrm{Im}\left[M_k(-j\omega)\right] d\omega$$

where $p = 3.14159$

$k = 1, 2$

The functions $G(\pi)$, $g(\pi)$ can then be expressed as :

$$G(\pi) = S_1(\pi) - S_2(\pi)$$

$$g(\pi) = \pi S_1(\pi) + (1 - \pi) S_2(\pi)$$

In the formula for $M_k(j\omega)$, π appears only in the first factor, $K(j\omega)$.

Hence, the computation of $S_k'(\pi)$ and therefore of $G'(\pi)$, involves one more integration, using $K'(j\omega)$ instead of $K(j\omega)$.

APPENDIX III

The matrix $R(\pi)$ has elements

$$R_{km}(\pi) = E\left\{ \left(g_k(\pi) - W_{n+1}^{k}\right)\left(g_m(\pi) - W_{n+1}^{m}\right) \,\middle|\, \pi \right\}$$

$$= g_k(\pi) \cdot g_m(\pi) - g_m(\pi)\; E\left(W_{n+1}^{k} \,\middle|\, \pi\right)$$

$$- g_k(\pi)\; E\left(W_{n+1}^{m} \middle| \pi\right) + E\left(W_{n+1}^{k}\; W_{n+1}^{m} \,\middle|\, \pi\right)$$

We have shown that

$$E\left(W_{n+1}^{k} \,\middle|\, \pi\right) = g_k(\pi)$$

Also,

$$W_{n+1}^{k} \cdot W_{n+1}^{m} = \begin{cases} W_{n+1}^{k} & \text{for} \quad k = m \\ 0 & \text{for} \quad k \neq m \end{cases}$$

Therefore,

$$R_{km}(\pi) = g_k(\pi)\left[\delta_{km} - g_m(\pi)\right]$$

where

$$\delta_{km} = \begin{cases} 1 & \text{for} \quad k = m \\ 0 & \text{for} \quad k \neq m \end{cases}$$

REFERENCES

[1] E.A. Patrick and J.C. Hancock : "Nonsupervised Sequential Classification and Recognition of Patterns," IEEE Trans. on IT, 1966, (3).

[2] S.J. Yakowitz : "Unsupervised Learning and the Identifiability of Finite Mixtures," IEEE Trans. on IT, 1970, (3).

[3] K. Fukunaga and W.L.G. Koontz : "A Criterion and an Algorithm for Grouping Data," IEEE Trnas. on C, 1970, (10).

[4] G. Nagy and G.L. Shelton : "Self Corrective Character Recognition System," IEEE Trans. on IT, 1966, (2).

[5] A.K. Agrawala : "Learning with a Probabilistic Teacher," IEEE Trans. on IT, 1970, (3).

[6] H.J. Scudder : "Adaptive Communication Receivers," IEEE Trans. on IT, 1965, (2).

[7] L.D. Davisson and S.C. Schwartz : "Analysis of a Decision Directed Receiver with Unknown Priors," IEEE Trans. on IT, 1970, (3).

[8] D. Kazakos : unpublished results on the study of a decision directed receiver with unknown mean and priors.

[9] J. Sacks : "Asymptotic Distribution of Stochastic Approximation Procedures," Ann. Math. Stat., 1958, (2)

[10] D. Sakrison : "Stochastic Approximation : A Recursive Method for Solving Regression Problems," Advances in Communication Systems, Vol. 2, A.V. Balakrishnan, ed. New York, Academic Press, 1966, pp. 51.

[11] K. Fukunaga and T. Krile : "Calculation of Bayes' Recognition Error for Two Multivariate Gaussian Distributions," IEEE Trans. on Comp, March, 1969.